Sampling Large Random Knots in a Confined Space

J. Arsuaga[1], T. Blackstone[2], Y. Diao[3], K. Hinson[3], E. Karadayi[4] and M. Saito[4]

[1] Department of Mathematics
San Francisco State University
1600 Holloway Ave.
San Francisco, CA 94132
[2] Department of Computer Science
San Francisco State University
1600 Holloway Ave.
San Francisco, CA 94132
[3] Department of Mathematics and Statistics
University of North Carolina at Charlotte
Charlotte, NC 28223
[4] Department of Mathematics
University of South Florida
4202 E. Fowler Avenue
Tampa, FL 33620

Abstract. DNA knots formed under extreme conditions of condensation, as in bacteriophage P4, are difficult to analyze experimentally and theoretically. In this paper, we propose to use the uniform random polygon model as a supplementary method to the existing methods for generating random knots in confinement. The uniform random polygon model allows us to sample knots with large crossing numbers and also to generate large diagrammatically prime knot diagrams. We show numerically that uniform random polygons sample knots with large minimum crossing numbers and certain complicated knot invariants (as those observed experimentally). We do this in terms of the knot determinants or colorings. Our numerical results suggest that the average determinant of a uniform random polygon of n vertices grows faster than $O(e^{n^2})$. We also investigate the complexity of prime knot diagrams. We show rigorously that the probability that a randomly selected 2-D uniform random polygon of n vertices is almost diagrammatically prime goes to one as n goes to infinity. Furthermore, the average number of crossings in such a diagram is at the order of $O(n^2)$. Therefore, the 2-dimensional uniform random polygons offer an effective way in sampling large (prime) knots, which can be useful in various applications.

AMS classification scheme numbers: 57M25

Submitted to: *J. Phys. A: Math. Gen.*

1. Introduction

Knotted circular polymers are commonly observed in chemistry and biology. Knots can be synthesized in chemistry laboratories for the purpose of studying chemical isomerism and chirality of molecules [23] and have been found in computational chemistry studies where long polyethylene chains in solution are simulated [40]. Biopolymers such as RNA, DNA and proteins also show a large repertoire of knots with distinct biological information. For instance, it has recently been found that the backbone of about 40 proteins in the protein data-bank are knotted (upon appropriate closure of the ends). These folding patterns provide a paradigm to study protein folding processes and are believed to have a functional role ([41], reviewed in [36]).

DNA knots are found as products of the action of diverse types of enzymes; such as recombinases [29, 39], topoisomerases [9, 18], polymerases [28] and condensins [20, 30], as well as as products of cyclization reactions of linear molecules [32, 34, 4]. Both the knotting probability and the knot complexity have been well characterized experimentally and computationally for the closure of DNA chains in free solution [34, 21, 32, 38]. Furthermore these studies have been supported by analytical results on polymer chains that estimate the rate of growth of the knotting probability and knot distribution as a function of the chain length [10, 15, 35].

The problem of DNA knotting in confinement remains largely unexplained despite its important biological implications in chromosome biology. A system particularly amenable for investigating this problem is that of bacteriophage P4 [22, 43]. The genome of these viruses are released in circular form when treated with protein denaturing agents. Analysis of these DNA circles has shown a very high knotting probability as well as extreme complexity of the knot populations [4]. Knowing the probability of occurrence of these knots as well as their complexity may help us to understand the organization of DNA inside bacteriophage. For example, analysis of the knot distribution of P4 knots of up to 7 crossings has revealed that the DNA is chirally organized while in the phage [5]. Furthermore, it has been conjectured that a good proportion of these knots are prime [2] as it is observed in simulation studies of collapsed polygons [6]. However due to experimental limitations DNA knots with more than 7 crossings have not yet been analyzed and remains a difficult challenge.

Several researchers have investigated the knotting probability as well as the knot complexity of polymer chains in confinement [4, 5, 19, 24, 25, 26, 37]. However progress in this area is thwarted by the lack of analytical results that support the computational efforts. Such results exist only for polygons in the simple cubic lattice that are confined to slabs or prisms [31] and remain open for polygons in confined volumes such as spheres or boxes.

Here we use the *uniform random polygon* (URP) model to investigate the complexity of knots confined to cubical boxes. The URP model was initially proposed by Millett [26] to study knotting in confined spaces and later used by Arsuaga et al. to investigate the linking probability of two polygons in confinement [3]. An obvious disadvantage of the URP model is that the flexibility of the chain is not well-defined except for limit cases. However, the URP model offers a valuable alternative investigating tool for us since it shows very similar topological qualitative results as other polymer models, and allows proving certain rigorous mathematical results while drastically reducing computational time.

This paper is organized as follows. We first derive some basic results concerning the mean average crossing number (mean ACN) of a uniform random polygon (Section 2). We find that the mean ACN of URP knots grows at the order $O(n^2)$. Such results give us some perspective about the complexity of the knot and of its projection diagrams. In Section 3, we use the determinant of a knot and the colorability of the knot to investigate the knotting probability and the complexity of the knots sampled. We find that the determinant increases at a rate of $O(e^{0.0014n^{2.4}})$, suggesting that it is possible that the rate of growth for the average determinant of an URP of n vertices is super exponential in general (i.e., faster than $O(e^n)$). For URPs with the number of vertices in the range of our numerical study, we found that the knotting probability fits well with the curve $1 - e^{-bn^3}$ where $b \approx 0.000082$, though this is not to be expected for large values of n. In section 4 we investigate the complexity of a population of knot diagrams and introduce the concept of almost diagrammatically prime diagrams. This study is motivated by the finding that prime knots may be prevalent over composite knots in confinement [2, 6] and because it is important in knot theory itself since prime knots in general are studied and tabulated through the study of their minimum diagrams which are diagrammatically prime. We rigorously prove that as n increases, the probability that a 2-D uniform random polygon of n vertices is almost diagrammatically prime goes to one at a rate of at least $1-O(\frac{1}{n^\nu})$ for some constant $\nu \geq 0.35$. We complement this analytical result by a numerical study in which we sample prime alternating knots using 2-D uniform random polygons as the knot diagrams. These numerical results show very complicated knots that can be achieved with as few as 12 vertices, supporting our argument that the 2-D URPs are good candidates for sampling (complicated) large prime knot diagrams. We end the paper by discussing possible applications of our results to the problem of DNA packing in bacteriophages and to the problem of generating complicated prime knots.

2. Uniform Random Polygons in a Confined Space

For $i = 1, 2, ..., n$, let $U_i = (u_{i1}, u_{i2}, u_{i3})$ be a three-dimensional random point that is uniformly distributed in the unit cube C^3 (or in a unit ball) such that U_1, U_2, ..., U_n are independent. Let e_i (called the i-th edge) be the line segment joining U_i and U_{i+1}, then the edges e_1, e_2, ..., e_n define a *uniform random polygon* R_n in the confined space (either the cube or the sphere), where e_n is the line segment joining U_n and U_1. If in this definition, we only consider the first two coordinates of each U_i, then we obtain a 2-dimensional uniform random polygon confined in the unit square C^2.

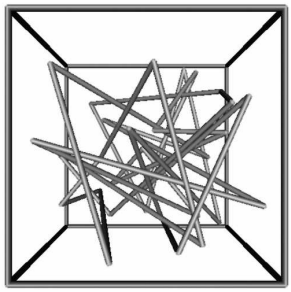

Figure 1. A uniform random polygon confined in the unit cube.

First, let us consider the case of two oriented random edges ℓ' and ℓ''. Since the end points of the edges are independent and are uniformly distributed in C^3, the probability that the projections of ℓ' and ℓ'' intersect each other is a positive number, which we will call $2p$. Notice that $2p$ is the same as the mean average crossing number between ℓ' and ℓ''.

Now consider a uniform random polygon R_n with n edges e_1, e_2, ..., e_n in that consecutive order. Let $a(e_i, e_j)$ be the average crossing number between e_i and e_j, then the average crossing number of R_n is

$$\chi_n = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i-1,i,i+1} a(e_i, e_j).$$

It follows that the expected value of the average crossing number of R_n is

$$E(\chi_n) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i-1,i,i+1} E(a(e_i, e_j)) = p(n-3)n.$$

Notice that $E(\chi_n)$ is identical to the mean average crossing number of a 2-D R_n in the case that the confining space is the unit square.

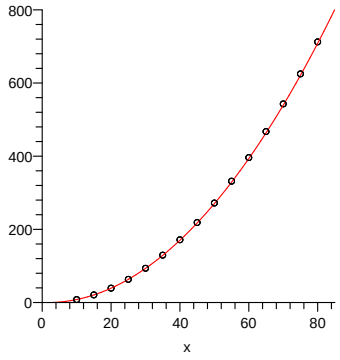

Figure 2. The mean ACN of uniform random polygons up to 80 vertices

Through computer simulations, we have estimated that $2p = 0.23 \pm 0.004$ so $E(\chi_n) \approx 0.115(n-3)n$. Figure 2 shows the mean ACN obtained through numerical simulations where the x-axis is the number of vertices, the y-axis is the mean ACN and the fitting curve is $y = 0.115(x-3)x$. Notice that the fit is near perfect.

It is worthwhile for us to point out here that the behavior of the mean ACN for URPs is very different from that of a random polygon without volume confinement. For example, the mean ACN of an equilateral random polygon of n vertices is shown to grow at a rate of $O(n \ln n)$ [11] and similar result hold for Gaussian random polygons as well [12].

3. The knot complexity of uniform random polygons

In this section, we wish to demonstrate the complexity of knots generated using the 3-D uniform random polygons in C^3. Determining the knot complexity of a random

polygon in a confined space is a difficult problem even when the number of vertices in the polygon is relatively small. For example, a uniform random polygon with 100 vertices typically has a regular projection with up to the order of 1115 crossings. Although some crossings can be eliminated by obvious Reidemeister I or II moves, it is not clear how many of these crossings can be removed in this fashion. It is likely that we may still have to deal with a diagram with hundreds of crossings. Computing a knot polynomial from such a diagram is out of question since we apparently lack the computing power at this stage. Instead, we will employ a relatively simple knot complexity measure that is easy to compute, namely the determinant and the coloring of knots. Let us first give the definition of the coloring of a knot for the convenience of our reader. Interested reader may refer to [27] for more details.

Let P be a projection of a knot K with m crossing points. Let $\mathcal{A} = \{A_1, \ldots, A_m\}$ be the set of over-arcs. A *coloring* (or a *p-coloring*) is a map $\mathrm{Col} : \mathcal{A} \to \mathbb{Z}_p$ for an odd prime p such that at every crossing the following condition is satisfied: If A_r and A_s are under-arcs and A_u is the over-arc, then $\mathrm{Col}(A_r) + \mathrm{Col}(A_s) \equiv 2\mathrm{Col}(A_u) \bmod p$.

The image $\mathrm{Col}(A_u)$ is called a *colour* assigned to the arc A_u. A coloring is called trivial if it assigns a single colour to all the arcs. The knot K is said to be p-colourable if there exists a non-trivial p-coloring.

Let K be a knot and let $\Delta_K(t)$ be the Alexander polynomial of K. $\Delta_K(-1)$ is called the *determinant* of K and is denoted by $\mathrm{Det}(K)$. It is known that a knot K is p-colourable for a prime p if and only if p divides $\mathrm{Det}(K)$ [16].

For a knot K, the *stick number* of K is defined as the minimum number of straight line segments required to form a polygonal representation of K. It is not hard to see that the stick number for any non-trivial knot is at least 6 [1]. In fact, only the trefoil can have stick number 6 [7]. Thus we will start from $n = 6$ in our numerical study. For each n, we generate m uniform random polygons of n vertices and compute the determinant for each one of them. We then compile our numerical result in a list $\mathrm{DET}(n, m)$ whose entries are integer pairs $[k, \ell]$, where k is a positive integer and ℓ is the frequency of uniform random polygons with determinant k. For example, $\mathrm{DET}(n, m) = \{[1, m]\}$ for all $n < 6$. For $n = 6$, our output was $\mathrm{DET}(6, 100000) = \{[1, 99534], [3, 466]\}$. This means that out of the 100000 uniform random polygons of 6 vertices we generated, 99534 of them are the unknots and 466 of them are trefoils. The following is a partial list of our numerical results concerning the determinants of the uniform random polygons, using only $n = 6$, 10, 12, 15, 17, 20. For clarity, we have omitted the entry for sample size since they are all 100000 in this study.

$$
\begin{aligned}
\mathrm{DET}(6) &= \{[1, 99534], [3, 466]\} \\
\mathrm{DET}(10) &= \{[1, 92336], [3, 6410], [5, 945], [7, 235], [9, 27], [11, 22], [13, 22], [15, 3]\} \\
\mathrm{DET}(12) &= \{[1, 85470], [3, 10749], [5, 2450], [7, 781], [9, 215], [11, 143], [13, 97], \\
&\quad [15, 37], [17, 14], [19, 12], [21, 13], [23, 4], [25, 1], [27, 2], [29, 4], [31, 1], \\
&\quad [35, 4], [39, 1], [47, 1], [53, 1]\} \\
\mathrm{DET}(15) &= \{[1, 72498], [3, 16415], [5, 5522], [7, 2417], [9, 987], [11, 682], [13, 491], \\
&\quad [15, 240], [17, 167], [19, 153], [21, 106], [23, 63], [25, 50], [27, 42], \\
&\quad [29, 20], [31, 29], [33, 18], [35, 20], [37, 13], [39, 7], [41, 9], [43, 9], \\
&\quad [45, 5], [47, 6], [49, 5], [51, 3], [53, 3], [55, 3], [57, 4], [59, 2], [69, 1], \\
&\quad [71, 2], [73, 2], [83, 1], [95, 1], [139, 1], [143, 1], [145, 1], [157, 1]\}
\end{aligned}
$$

DET(17) = {[1, 63211], [3, 18514], [5, 7632], [7, 3742], [9, 2012], [11, 1267], [13, 909], [15, 555], [17, 422], [19, 330], [21, 284], [23, 157], [25, 130], [27, 134], [29, 81], [31, 85], [33, 65], [35, 54], [37, 51], [39, 42], [41, 21], [43, 34], [45, 37], [47, 18], [49, 24], [51, 12], [53, 15], [55, 15], [57, 12], [59, 12], [61, 18], [63, 3], [65, 6], [67, 9], [69, 6], [71, 5], [73, 2], [75, 6], [77, 5], [79, 4], [81, 3], [83, 4], [85, 2], [87, 1], [89, 2], [91, 2], [93, 4], [95, 3], [97, 2], [99, 4], [101, 1], [103, 3], [107, 2], [109, 2], [111, 2], [113, 1], [115, 2], [119, 2], [121, 1], [125, 2], [131, 1], [139, 1], [145, 2], [165, 1], [167, 1], [173, 1], [181, 2], [185, 1], [203, 1], [207, 1], [293, 1]}

DET(20) = {[1, 48628], [3, 19801], [5, 9452], [7, 5589], [9, 3642], [11, 2445], [13, 1938], [15, 1303], [17, 954], [19, 910], [21, 698], [23, 480], [25, 454], [27, 480], [29, 340], [31, 286], [33, 216], [35, 215], [37, 181], [39, 165], [41, 155], [43, 114], [45, 122], [47, 90], [49, 87], [51, 82], [53, 83], [55, 82], [57, 68], [59, 61], [61, 54], [63, 60], [65, 39], [67, 33], [69, 37], [71, 35], [73, 26], [75, 40], [77, 19], [79, 24], [81, 30], [83, 24], [85, 19], [87, 23], [89, 16], [91, 18], [93, 21], [95, 15], [97, 11], [99, 19], [101, 13], [103, 11], [105, 16], [107, 9], [109, 6], [111, 4], [113, 12], [115, 13], [117, 13], [119, 5], [121, 9], [123, 8], [125, 8], [127, 6], [129, 8], [131, 9], [133, 12], [135, 3], [137, 1], [139, 5], [141, 2], [143, 7], [145, 3], [147, 2], [149, 2], [151, 4], [153, 7], [155, 2], [157, 3], [159, 2], [161, 1], [163, 2], [165, 3], [167, 4], [169, 2], [171, 3], [173, 2], [175, 1], [177, 4], [179, 4], [181, 2], [183, 2], [185, 1], [187, 4], [189, 2], [191, 2], [193, 3], [195, 1], [199, 1], [201, 3], [203, 1], [207, 2], [209, 2], [211, 3], [213, 1], [215, 4], [217, 1], [219, 1], [223, 3], [225, 2], [229, 1], [233, 1], [235, 2], [237, 1], [241, 1], [243, 1], [245, 1], [249, 2], [257, 2], [259, 1], [265, 1], [269, 1], [281, 1], [289, 1], [293, 1], [317, 1], [319, 1], [325, 1], [331, 1], [341, 1], [369, 1], [381, 1], [389, 1], [391, 1], [399, 1], [403, 1], [409, 1], [419, 1], [435, 1], [451, 1], [455, 1], [459, 1], [475, 1], [483, 1], [495, 1], [513, 1], [519, 1], [521, 1], [567, 1], [771, 1], [807, 1]}

First, we point out that the numerical data can be used to give a lower bound on the knotting probability of a 3-D URP. Figure 3 is the plot of the percentage of URPs with non-trivial determinants, where the horizonal axis is the number of vertices of the URPs. The fitted curve given here is $1 - \exp(-0.000082n^3)$. But this is not to be expected as a general rule as the following argument shows that the trivial knot probability of an R_n is at least of order $\exp(-n \ln n)$, which is larger than $\exp(-0.000082n^3)$ for large values of n. Let us divide the unit square C^2 (which is the base of the unit cube and is on the xy plane) into k^2 equal size squares such that $(k-1)^2 < n \leq k^2$. Number these small squares in a sequential order as shown in Figure 4 (for the case of $k = 4$). Apparently, for each given $1 \leq j \leq n$, the probability for the projection of U_j (to the xy plane) to fall in the j-th square is equal to the area of the square, which is $1/k^2$. Thus the probability this happens for all j at the same time is precisely $(1/k^2)^n$, which is of the order $(1/n)^n = n^{-n} = \exp(-n \ln n)$, as one can easily check. However when this happens, the polygon is the trivial knot. On the other hand, since the maximum number of crossings in the projection of a URP of n segments is bounded above by n^2, this numerical result may suggest that the trivial knot probability of a URP tends to 0 faster than $e^{-b_0 c}$ where b_0 is some constant and c is the number of crossings in the projection of the R_n onto the xy plane.

Second, as n increases, the maximum determinant in our sample increases

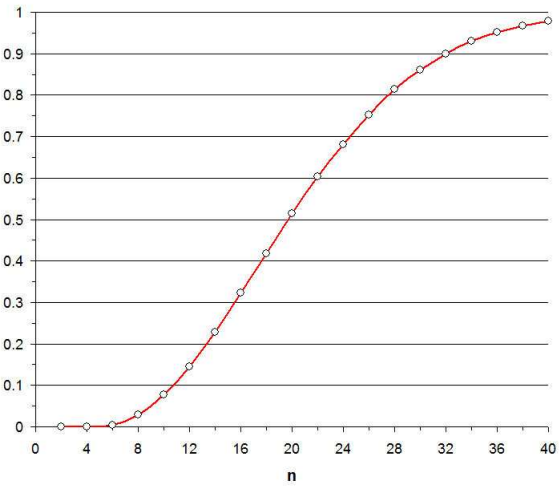

Figure 3. The lower bound of knotting probability for URPs up to 40 segments based on the percentage of URPs with non-trivial determinants.

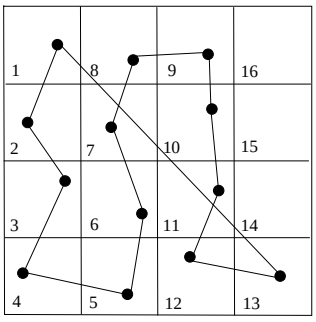

Figure 4. An R_n is a trivial knot if its projection follows the pattern shown here. This is the case of $k = 4$ and $n = 13$.

dramatically. For example, for $n = 59$, a small sample with only 50 data points yields a maximum determinant of 12088427779. The following figure is the logarithmic plot of the average determinant from our samples. In fact, the logarithmic plot of the average determinants given in Figure 5 suggests that the growth of the average determinant of R_n is likely to be faster than $O(\exp(n^2))$.

Third, we present our numerical output on the coloring of the uniform random knots. Let n be the number of vertices of the random polygon, m be the number of uniform random polygons generated in our data set, and we will let $\mathrm{COL}(n, m)$ be the list of outputs, which consists of entries of the form $[p, \ell]$, where the first entry is of the form $[1, \ell]$ with ℓ indicating the number of polygons that can only be trivially colored and $p > 2$ is a prime in other entries with the corresponding ℓ the number of polygons in the sample that are p-colourable. For example, $\mathrm{COL}(10, 1000) = \{[1, 914], [3, 68], [5, 21]\}$ means that out of the 1000 uniform random

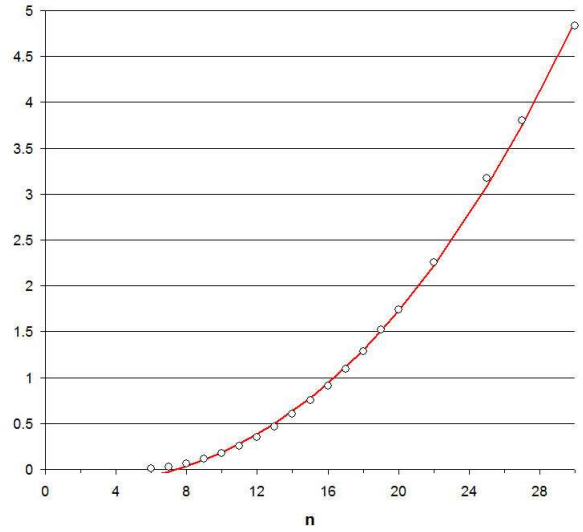

Figure 5. The logarithmic plot of the average determinants of uniform random polygons with up to 30 vertices, where the equation of the fitting curve is $y = -.18 + 0.0014x^{2.4}$

polygons generated, 914 of them can only be trivially colored, 68 of them have determinants divisible by 3 (hence are 3-colourable) and 21 of them have determinants divisible by 5 (hence are 5-colourable). Notice that since a knot can be both 3-colourable and 5-colourable, it is possible that a polygon is multiply counted in this method. For example, in this example of $\mathrm{COL}(10, 1000) = \{[1, 914], [3, 68], [5, 21]\}$, there are three polygons that are both 3-colourable and 5-colourable. That is, the summation of the ℓ entries of $\mathrm{COL}(n, m)$ is generally larger than m. For each n between 6 and 30, we generated 100000 uniform random knots. A few selected output list are presented in the following table. Again we omit the sample size in COL since they are all 100000.

$$
\begin{aligned}
\mathrm{COL}(6) &= \{[1, 99534], [3, 466]\} \\
\mathrm{COL}(10) &= \{[1, 92336], [3, 6440], [5, 948], [7, 235], [11, 22], [13, 22]\} \\
\mathrm{COL}(12) &= \{[1, 85470], [3, 11017], [5, 2492], [7, 798], [11, 143], [13, 98], [17, 14], \\
&\quad [19, 12], [23, 4], [29, 4], [31, 1], [47, 1], [53, 1]\} \\
\mathrm{COL}(15) &= \{[1, 72498], [3, 17828], [5, 5842], [7, 2548], [11, 704], [13, 499], [17, 499], \\
&\quad [19, 158], [23, 64], [29, 21], [31, 29], [37, 13], [41, 9], [43, 9], \\
&\quad [47, 6], [53, 3], [59, 2], [71, 2], [73, 2], [83, 1], [139, 1], [157, 1]\} \\
\mathrm{COL}(17) &= \{[1, 63211], [3, 21698], [5, 8448], [7, 4117], [11, 1358], [13, 959], \\
&\quad [17, 438], [19, 345], [23, 166], [29, 85], [31, 89], [37, 54], \\
&\quad [41, 21], [43, 34], [47, 18], [53, 15], [59, 12], [61, 18], [67, 9], \\
&\quad [71, 5], [73, 2], [79, 4], [83, 4], [89, 2], [97, 2], [101, 1], [103, 3], \\
&\quad [107, 2], [109, 2], [113, 1], [131, 1], [139, 1], [167, 1], [173, 1], \\
&\quad [181, 2], [293, 1]\}
\end{aligned}
$$

$$\begin{aligned}\mathrm{COL}(20) \;=\; \{&[1,48628],[3,26909],[5,11807],[7,6732],[11,2810],[13,2186],\\ &[17,1074],[19,1013],[23,535],[29,369],[31,312],[37,187],\\ &[41,165],[43,126],[47,94],[53,86],[59,65],[61,56],[67,36],\\ &[71,36],[73,27],[79,25],[83,26],[89,16],[97,11],[101,13],\\ &[103,11],[107,9],[109,6],[113,12],[127,7],[131,9],[137,1],\\ &[139,5],[149,2],[151,4],[157,3],[163,2],[167,4],[173,3],\\ &[179,4],[181,2],[191,2],[193,3],[199,1],[211,3],[223,3],\\ &[229,1],[233,1],[241,1],[257,2],[269,1],[281,1],[293,1],\\ &[317,1],[331,1],[389,1],[409,1],[419,1],[521,1]\}\end{aligned}$$

It is evident that even for the relatively small value of $n = 20$, many polygons sampled are fairly complicated knots. On the other hand, the average number of colours per random polygon turns out to be a fairly slow growth number. For the data set we listed above for $n = 6$, 10, 12, 15, 17 and 20, the corresponding average numbers of colours per polygon are 1.00466, 1.07667, 1.14585, 1.27912, 1.37919 and 1.54826, indicating that for most polygons in the small range of n we have studied can only allow one coloring, trivial or non-trivial. On the other hand, the average value of the maximum coloring an URP admits can grow at much faster rate. For example, in our small sample of 50 URPs of $n = 59$ vertices, we already had a maximum determinant of 12088427779, however it has only 4 nontrivial prime factors (17, 19, 53 and 706141). Further studies are needed in order to have a more complete picture. However, the factoring of many large determinants is very time consuming in computation and is the main obstacle in carrying out this further study.

4. Uniform random polygons as knot diagrams

In this section, we are primarily interested in producing diagrammatically prime diagrams, that is, 4-regular graphs without cut vertices and are more than 2-edge connected. A cut vertex of a graph is a vertex of the graph so that the removal of the vertex (together with the edges incident to it) will divide the graph into at least two disjoint components. A graph is called k-edge connected if the graph is connected and removing any $k-1$ edges of the graph will not change that. For a plane graph that is also a knot diagram (not a link diagram), if removing any two edges cannot disconnect the graph, then removing any three edges will not disconnect the graph either. That is, if the graph is 3-edge connected, then it is also 4-edge connected [13]. Here we will treat the 2-D uniform random polygons as random knot diagrams. Notice that when we treat a 2-D polygon as a graph, the vertices of the graph are the crossings of the polygon while the vertices of the polygon itself are not vertices of the graph. Since no crossing may involve three or more distinct segments of the polygon with probability one, we will obtain a 4-regular plane graph when we sample a 2-D uniform random polygon with probability one. However there is no automatic guarantee that such a graph contains no cut vertices and is 4-edge connected. Finally, we point out that in the case that the graph contains a loop vertex, then it is 2-edge connected (however a loop vertex is not a cut vertex).

Definition 1 *A plane 4-regular graph is said to be almost diagrammatically prime if it contains no cut vertices and becomes 4-edge connected once the loop edges (if there are any) are removed (and the corresponding vertices are no longer treated as vertices) (see Figure 6).*

From an almost diagrammatically prime diagram, one can easily construct an alternating knot that is also prime, as shown in Figure 7.

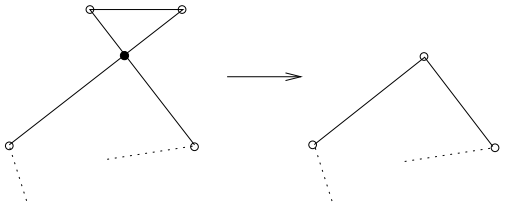

Figure 6. The removal of a loop edge in plane graph produced by a 2-D polygon: The solid dot is the vertex of the graph which is a crossing of the polygon whereas the vertices of the polygon are marked by circles.

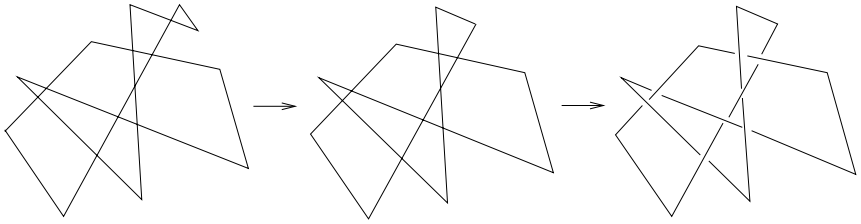

Figure 7. The construction of an alternating prime knot from an almost diagrammatically prime diagram.

The following is our main result of this section.

Theorem 1 *Let R_n be a 2-D uniform random polygon confined in C^2. Then as a plane 4-regular graph, the probability that R_n is almost diagrammatically prime is of the order at least $1 - O(\frac{1}{n^\nu})$ where the constant ν can be chosen to be at least* 0.35.

We need some preparation before we proceed to prove the theorem. Let α be a constant number between -1 and 0. Its exact value will be determined later. Let us first consider a line segment ℓ confined in C^2 whose length is at least n^α and at least one end point of it is of a distance n^α away from the boundary of C^2. Assume further that there is a random line segment ℓ' whose two end points are independent and are uniformly distributed in C^2. We would like to bound from below the probability that ℓ' intersects ℓ. We claim that this probability is at least of the order $O(n^{2\alpha+\beta})$ where β is some negative constant between α and 0. (We will eventually use $\alpha = -0.45$ and $\beta = -0.05$.) To see this, let us extend ℓ to a full straight line L. L divides C^2 into two parts. Choose the part that has a bigger area and draw a line L' through this part that is parallel to L and is of a distance n^β to L. If n is large, n^β is small so the area of the part of C^2 divided by L' that does not contain ℓ is at least $1/4$. It follows that the probability for one end P of ℓ' to fall in this area is at least $1/4$. When this happens, if the other end Q of ℓ' falls into the shaded region as shown in Figure 8, ℓ' will intersect ℓ. An extreme case for the position of P is shown in the figure (marked as P'). It is easy to see that the shaded region contains the right triangle (marked in black) and the height of the right triangle is of the order $O(n^{\alpha+\beta})$. It follows that the

area of the shaded region is of an order at least $O(n^{2\alpha+\beta})$. Thus the probability of ℓ' intersecting ℓ is at least $O(n^{2\alpha+\beta}) \cdot \frac{1}{4} = O(n^{2\alpha+\beta})$, since P and Q are independent random points. This leads to the following lemma.

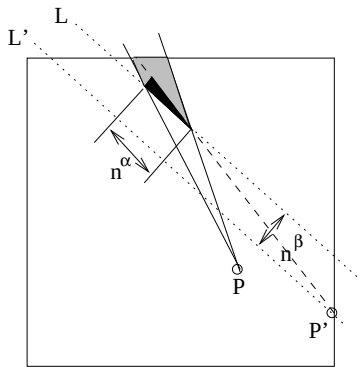

Figure 8. The shaded region has an area at least of the order $O(n^{\alpha+\beta})$.

Lemma 1 *Let α and β be two negative constants such that $\alpha < \beta$, $2\alpha + \beta + 1 > 0$ and let ℓ be a line segment confined in C^2 whose length is at least n^α. Furthermore, at least one end point of ℓ is of a distance n^α away from the boundary of C^2. Let ℓ_1, ℓ_2, ..., ℓ_k be k random line segments whose end points are independent random points uniformly distributed in C^2. Then if $k \geq an$ for some constant $a > 0$, then the probability that at least one ℓ_j intersects ℓ is of the order at least $1 - \exp(-bn^{1+2\alpha+\beta})$ for some constant $b > 0$.*

Proof. Since the end points of the ℓ_j's are independent, the probability that none of the ℓ_j's intersects ℓ is at most $O((1 - n^{2\alpha+\beta})^k)$. Substituting an for k, we have

$$\begin{aligned}(1 - n^{2\alpha+\beta})^{an} &= \left((1 - n^{2\alpha+\beta})^{n^{-2\alpha-\beta}}\right)^{an^{1+2\alpha+\beta}} \\ &< \exp(-bn^{1+2\alpha+\beta})\end{aligned}$$

for some constant $b > 0$ since

$$\lim_{n\to\infty} (1 - n^{2\alpha+\beta})^{n^{-2\alpha-\beta}} = e^{-1}.$$

We are now ready to prove our theorem.

Proof. Notice that if the diagram has a cut vertex, then the diagram is also at most 2-edge connected since removing two edges at the same side of the vertex will also disconnect the graph in such a case. Therefore, we need only to consider the case when the diagram is at most 2-edge connected. Let A_j $(1 \leq j \leq n/2 - 1)$ denote the following event: there exists a pair of segments ℓ_1 and ℓ_2 in R_n such that the removal of ℓ_1 and ℓ_2 in R_n will result in two non-overlapping random walks, one is of length j. Assume that the diagram becomes disconnected after removing two edges e_1 and e_2. Since e_1 and e_2 belong to some segments ℓ_1 and ℓ_2 of the polygon R_n, deleting ℓ_1 and ℓ_2 from R_n results in two non-overlapping random walks S_1 and S_2 and the shorter one of S_1 and S_2 will contain j segments for some $j \leq n/2 - 1$. In other word, if R_n

produces a diagram that is at most 2-edge connected, then one of the A_j will happen. Let us concentrate on the shorter one, say it is S_1. In the cases of A_1 and A_2, the most we can get from the shorter component after the removing of the two segments is a simple loop. Thus the probability of getting an almost diagrammatically prime graph is bounded below by $1 - \sum_{3\leq j\leq n/2-1} P(A_j)$. Let us consider the case of A_3. Let ℓ_1 and ℓ_2 be two segments of R_n separated by three consecutive segments that do not overlap with the other segments of R_n. Say these three segments have end points U_1, U_2, U_3 and U_4 (these are independent 2-D random points uniformly distributed in C^2). If one of the segments in S_1 satisfies the conditions of ℓ in Lemma 1, then the probability for some segments of S_2 to intersect this segment is of the order at least $1 - \exp(-bn^{1+2\alpha+\beta})$. That is, the probability that S_1 and S_2 are non-overlapping is at most $\exp(-bn^{1+2\alpha+\beta})$ in this case. So what is the probability that none of the three segments satisfies the conditions of ℓ in Lemma 1? Let B be the set of points of C^2 that are within a distance n^α from the boundary of C^2. There are five cases to consider depending on how many of of U_1, U_2, U_3 and U_4 fall in B.

Case 0. None of U_1, U_2, U_3 and U_4 falls in B. This actually happens with a large probability since the area of B is small. But this means that the four vertices will have to be close to each other. That is, once we have chosen U_1, then U_2 can only be in the ball of radius n^α centered at U_1, and U_3 has to be in the ball of radius n^α centered at U_2, and so on. The probability of this is apparently of the order of $O(n^{3\alpha})$.

Case 1. One of U_1, U_2, U_3 and U_4 falls in B. This happens with probability of order n^α since the area of B is of that order. Say U_2 is in B. But then U_1 and U_3 will have to be close to U_2 and U_4 will have to be close to U_3. An argument similar to the above then yields a probability of the order $O(n^{4\alpha})$.

Case 2. Two of U_1, U_2, U_3 and U_4 fall in B. This happens with probability of order $n^{2\alpha}$. Say U_1 and U_2 are in B. But then U_3 will have to be close to U_2 and U_4 will have to be close to U_3. We again obtain a probability of the order $O(n^{4\alpha})$.

Cases 3 and 4. Three or four of U_1, U_2, U_3 and U_4 fall in B. These are similar to case 2 and the probability in both cases is of the order $O(n^{4\alpha})$.

Summarizing the above, we see that the probability for S_1 and S_2 to be non-overlapping is at most of the order $\exp(-bn^{1+2\alpha+\beta}) + O(n^{3\alpha})$. Since this calculation is based on a particular choice of the end points of the three segments in S_1 and there are $O(n)$ different such choices, the probability $P(A_3)$ is at most of the order $n\exp(-bn^{1+2\alpha+\beta}) + O(n^{1+3\alpha})$.

In general, we have $P(A_j) \leq n\exp(-bn^{1+2\alpha+\beta}) + O(n^{1+j\alpha})$ by using an argument similar to the above. This leads to

$$\sum_{3\leq j\leq n/2-1} P(A_j) \leq n^2\exp(-bn^{1+2\alpha+\beta}) + nO(\frac{n^{3\alpha}}{1-n^\alpha}) = O(n^{1+3\alpha}),$$

since $n^2\exp(-bn^{1+2\alpha+\beta}) < O(n^{1+3\alpha})$ and $nO(\frac{n^{3\alpha}}{1-n^\alpha}) = O(n^{1+3\alpha})$. Clearly, if we choose $\alpha = -0.45$ and $\beta = -0.05$, then for large values of n, the probability that R_n produces an almost diagrammatically prime 4-regular plane graph is at least $1 - O(n^{-0.35})$.

Theorem 1 guarantees that a 2-D URP of n vertices is (almost surely) an almost diagrammatically prime 4-regular plane graph when n is large. However, how many

n	ACN	AL	RACN
10	8.047	1.234	6.813
15	20.824	0.831	19.993
20	39.296	0.626	38.670
25	63.319	0.531	62.788
30	93.580	0.476	93.104
35	129.476	0.425	129.051
40	171.4934	0.385	171.109
45	218.689	0.365	218.324
50	271.966	0.335	271.631
55	331.693	0.313	331.380
60	396.174	0.290	395.884
65	467.321	0.288	467.033
70	542.954	0.270	542.684
75	624.979	0.259	624.720

Table 1. ACN is the average number of crossings, AL is the average number of loop edges (namely the crossings that can be removed via Reidemeister type I moves) and RACN is the reduced average number of crossings (namely ACN − AL).

loop edges are there in such a plane graph on the average? This is a valid question since too many such edges may affect the total number of non-trivial crossings in the diagram. Also, what if n is relatively small? To answer this question, we generated 100000 2-D URPs for each n ranging from 10 to 75 (with an increment of 5) and computed the average crossings before and after the loop edges are removed from the diagrams. The results are listed in Table 1. It is clear from the table that the loop edges occur very seldom even for small values of n.

Finally, we wish to demonstrate the complexity of the alternating knots generated using the method described in the last section. For comparison purposes, we will again compute the determinants of the generated URPs. However, it turns out that these polygons are much, much more complicated and the computation time increased dramatically. We thus only focused our effort for one single value of n, namely $n = 12$. We generated 1000 2-D URPs and converted them into alternating knots according to the procedure described in the last section, and computed their determinants. Out of these 1000 knots, not a single one allows only the trivial coloring. The following is a partial list of the results.

$$\begin{aligned}\mathrm{DET}(12,1000) \quad = \quad & \{[1,0],[3,38],[5,38],[7,32],[9,13],[11,27],[13,7],[15,25],\\ & [17,15],[19,20],[21,10],[23,10],[25,10],[27,7],[29,20],\\ & [31,10],[33,10],[35,3],[37,5],[39,4],[41,14],[43,5],\\ & [45,7],[47,1],[49,5],[51,12],[53,4],[55,5],[57,4],[59,4],\\ & [61,4],[63,4],[65,5],[67,8],[69,4],[71,6],[73,1],[75,3],\\ & [79,1],[83,1],\cdots,[868029,1],[1728883,1],[15445771,1]\}\end{aligned}$$

Notice that the first missing odd number (other than 1) is 77 from this list and the distribution of the frequency is not strictly decreasing as the determinant value increases. The average determinant here is 41889.48. On the other hand, for the 3-D

URPs with 12 vertices, the following results are from a comparable sample of the same sample size 1000. In this case, the average determinant is merely 1.42.

$$\mathrm{DET}(12, 1000) \quad = \quad \{[1, 851], [3, 116], [5, 21], [7, 5], [9, 4], [11, 1], [13, 1], [21, 1]\}$$

These numerical results are indeed very convincing that the 2-D URPs lend us an efficient method in generating large prime knots.

5. Conclusions and ending remarks

Circular polymeric chains form complicated knots with high probability when confined to small volumes [4, 24, 25, 26, 37]. This fact is particularly relevant in chromosome biology since the DNA molecule needs to be highly condensed to fit in the cell nucleus [17]. Here we have investigated properties of knots generated by the URP model since this model may be informative for cases of extreme polymer condensation as it is the case of DNA molecules packed in bacteriophages, in some animal viruses [8] and in DNA-lipo complexes [33].

As mentioned before, the main disadvantage of the URP model is that the flexibility of the chain is not properly defined. However it is still possible to use this model to study physical systems. For instance, since the knotting probability for DNA circles extracted from phage P4 is about 0.95 and because of the results in Figure 3 one may model such system with 32 segments. From such a model one obtains the following results.

1) Average projections show around 107 crossings (see Figure 2) and many of these crossings can be removed via Reidemeister moves I and II. In fact, a sample of 100,000 URPs with 32 vertices yields a mean of 107.338 crossings per projection. After simple Reidemeister moves of type I and II are applied, the mean number of crossings per projection is reduced to 64.729. This number is somewhat larger than the ACN estimated experimentally. However the ACN of the knots found in phage capsids was estimated using the linear relationship between knot complexity and gel migration proposed in [42] and the actual complexity of these knots is still unknown.

2) The knot distributions that the URP model generates, originally studied by Millett in [26], are also consistent with those found using other polymer models to study DNA packing in phages [5, 25].

3) Knots formed under conditions of confinement are very complex and it is predicted that the average determinant of the knots can reach values of $O(\exp(n^2))$.

4) Since it is believed that knots formed under confinement are mostly prime [2, 6] and since distinguishing between prime and composite knots is a difficult task, we have investigated sets of knot diagrams that are diagrammatically prime. We have rigorously shown that a 2-D URP with n vertices is almost diagrammatically prime with a probability at least $1 - O(n^{-0.35})$. Our numerical study on this subject shows that most 2-D URPs with n vertices are indeed diagrammatically prime as n increases.

We believe that this method of generating large diagrammatically prime knots may be useful in knot theory and in polymer physics. Furthermore we expect

that similar analytical techniques as those presented here may be of help to show rigorous results concerning the growth rate of the knotting probability for polymers in confinement. Finally, we would like to make a few remarks about the computational aspects of this work.

Generating random knots is a task we have to perform both in applied and theoretical studies. From a pure mathematical point of view, an ideal random knot generator would select a knot with a given minimum crossing number uniformly from the set of all knots with that minimum crossing number. Of course, random knots generated this way are not to be confused with the various random polygons used to model subjects such as ring polymers (whose projections are usually highly redundant in crossings and are usually composite knots). However, very little is known about the space of large knots (that is, knots with large minimum crossing numbers). Consequently, it is not possible to develop such an ideal algorithm at this time. Thus, the best one can hope for is a large knot generating method with no obvious bias towards certain classes of knots that is easy to implement with a short runtime.

The proposed method in the last section seems to satisfy these naive criteria (at least for the alternating knots): the polygons are sampled with equal probability, they are very easy to generate and most of them are indeed large knots. The advantage of this method over the other known methods [14] is that it is much easier and quicker to produce large diagrammatically prime knot diagrams (and consequently prime alternating knots) with this method. Furthermore, this method is obviously ergodic, meaning that any prime (alternating) knot can be generated by this method with a positive probability. However, it does share a common drawback with the other two methods [14]: one cannot pre-determine the exact minimum crossing number of the knot or the knot diagram to be sampled. One way to amend this problem is through over-sampling. For example, if we want to generate 1000 prime alternating knots with minimum crossing number 20. Then we can simply generate many more diagrammatically prime diagrams (with, say, n up to 40) and simply take the first 1000 that has 20 crossings. This is practical since it is very easy to generate these polygons so the total runtime will not be too large.

Acknowledgement. This work was supported in part by a NIH grant 2S06GM52588-12 and a grant from the SFSU Center for Computing in the Life Sciences to J. Arsuaga and by NSF grants DMS-0301089, DMS-0603876 to M. Saito.

References

[1] Adams C 1994, *The Knot Book*, W.H. Freeman and Company, New York.
[2] Arsuaga J 2000, PhD dissertation, The Florida State University.
[3] Arsuaga J, Blackstone T, Diao Y, Karadayi E and Saito M 2007, *J. Physics A* **40** 1925–36.
[4] Arsuaga J, Vazquez M, Trigueros S, Sumners D W and Roca J 2002, *Proc. Natl. Acad. Sci. USA* **99** 5373–7.
[5] Arsuaga J, Vazquez M, McGuirk P, Trigueros S, Sumners D W and Roca J 2005, *Proc. Natl. Acad. Sci. USA* **102** 9165–9.
[6] Baiesi M, Orlandini E and Stella A L 2007, preprint.
[7] Calvo J and Millett K 1998, *Ideal Knots,* Ser. Knots Everything **19**, World Scientific 107–28.
[8] Casjens S 1997, *Structural Biology of Viruses*, Oxford Univ. Press 3–37.
[9] Dean F B, Stasiak A, Koller T and Cozzarelli N R 1985, *J. Biol. Chem.* **260**(8), 4975–83.
[10] Diao Y 1995, *J. Knot Theory Ramifications* **4**(2) 189–96.
[11] Diao Y, Dobay A, Kusner R, Millett K and Stasiak A 2003, *J. Physics A* **36**(46), 11561-74.

[12] Diao Y and Ernst C 2005, *Physical and numerical models in knot theory*, Ser. Knots Everything **36**, World Scientific 275–92.
[13] Diao Y, Ernst C and Yu X 2004, *Topology Appl.* **136**(1) 7-36.
[14] Diao Y, Ernst C and Ziegler U 2005, *Physical and numerical models in knot theory*, Ser. Knots Everything **36**, World Scientific 473–94.
[15] Diao Y, Pippenger N and Sumners D W 1994, *J. Knot Theory Ramifications* **3**(3) 419–29.
[16] Fox R H 1962, *Topology of 3-manifolds and related topics*, Prentice-Hall, 120–67.
[17] Holmes V and Cozzarelli N R 2000, *Proc. Natl. Acad. Sci. USA* **97** 1322–4.
[18] Hsieh T 1983, *J. Biol. Chem.* **258**(13) 8413–20.
[19] Hua X, Baghavan B, Nguyen D, Arsuaga J and Vazquez M 2007, *Topology Appl.* **154** 1381–97.
[20] Kimura K, Rybenkov V V, Crisona N J, Hirano T and Cozzarelli N R 1999, *Cell* **98** 239–48.
[21] Klenin K V, Vologodskii A V, Anshelevich V V, Dykhne A M and Frank- Kamenetskii M D 1988, *J. Biomol. Struct. Dyn.* **5** 1173–85.
[22] Liu L F, Davis J L and Calendar R 1981, *Nucleic Acids Res.* **9** 3979–89.
[23] Lukin O and Vogtle F 2005, *Ange. Chem. Int. Ed.* **44** 1456–77.
[24] Mansfield M L 1994, *Macromolecules* **27** 5924–6.
[25] Micheletti C, Marenduzzo D, Orlandini E and Sumners D W 2006, *J. Chem. Phys.* **124** 64903.1 –.10.
[26] Millett K 2000, *Knots in Hellas'98 (Delphi)*, Ser. Knots Everything **24**, World Scientific 306–34.
[27] Murasugi K 1996, *Knot Theory and its Applications*, Birkhäuser, Boston.
[28] Olavarrieta L, Martinez-Robles M L, Hernandez P, Krimer D B and Schvartzman J B 2002, *Mol. Microbiol.* **46** 699–707.
[29] Pathania S, Jayaram M and Harshey R M 2002, *Cell* **109** 425–36.
[30] Petrushenko Z M, Lai C H, Rai R and Rybenkov V V 2006, *J. Biol. Chem.* **281**(8) 4606–15.
[31] Janse van Rensburg E J, Orlandini E and Whittington S G 2006, *J. Physics A* **39** 13869-903.
[32] Rybenkov V V, Cozzarelli N R and Vologodskii A V 1993, *Proc. Natl. Acad. Sci. USA* **90**(11) 5307–11.
[33] Schmutz M, Durand D, Debin A, Palvadeau Y, Eitienne E R and Thierry A R 1999, *Proc. Natl. Acad. Sci. USA* **96**, 12293–8.
[34] Shaw S Y and Wang J C 1993, *Science* **260** 533–6.
[35] Sumners D W and Whittington S G 1989, *J. Physics A* **22**(13) 1471–4.
[36] Taylor W R 2007, *Comp. Biol. Chem.* **31** 151-62.
[37] Tesi M C, Jance van Rensberg E J, Orlandini E and Whittington S G 1994, *J. Physics A* **27** 347–60.
[38] Tesi M C, Janse van Rensburg E J, Orlandini E, Sumners D W and Whittington S G 1994, *Phys. Rev. E.* **49** 868–72.
[39] Vazquez M, Colloms S and Sumners D W 2005, *J. Mol. Biol.* **346** 493–504.
[40] Virnau P, Kantor Y and Kardar M 2005, *J. Am. Chem. Soc.* **127** 15102–6.
[41] Virnau P, Mirny A L and Kardar M 2006, *PLoS. Comput. Biol.* **152** 1074–9.
[42] Vologodskii A, Crisona N J, Laurie B, Pieranski P, Katritch V, Dubochet J and Stasiak A 1998, *J. Mol. Biol.* **278** 1–3.
[43] Wolfson J S, McHugh G L, Hooper D C and Swartz M N 1985, *Nucleic Acids Res.* **13** 6695–702.